it's owl

ARBEITSWELT PLUS

DAS "BLACK-BOX-PHÄNOMEN" IN DER KI-ENTWICKLUNG

Methodische Ansätze zur Schaffung von Transparenz und der Verbesserung des Zusammenspiels von Mensch, Technik und Organisation

Alexander Kuhn
Stefan Hartmann

WORKING PAPER #8

inIT | TH OWL
Fraunhofer IEM

ÜBER DAS KOMPETENZZENTRUM ARBEITSWELT.PLUS

Wie wird Künstliche Intelligenz die Arbeitswelt verändern? Wie gelingt es, Veränderungen der Arbeitswelt gemeinsam zu gestalten? Und wie können Beschäftigte auf den Wandel eigentlich vorbereitet werden? Antworten auf diese Fragen liefern wir als Kompetenzzentrum Arbeitswelt.Plus.

Unserem gemeinsamen Leitmotiv **Mensch. Industrie. Morgen.** entsprechend entwickeln Hochschulen und Unternehmen aus OstWestfalenLippe im Kompetenzzentrum gemeinsam mit der IG Metall Ansätze für die Einführung von Künstlicher Intelligenz in der Arbeitswelt, beispielsweise im Hinblick auf die Arbeitsplatzgestaltung und die Qualifizierung von Mitarbeiter:innen.

ÜBER DIE WORKING-PAPER-REIHE

Damit die Ausprägung der künftigen Arbeitswelt nicht allein technologisch geprägt wird, braucht es eine **ganzheitliche Gestaltung**. Deshalb führt das Kompetenzzentrum Arbeitswelt.Plus Erkenntnisse der Arbeitsforschung im Kontext von KI-Anwendungen zusammen und entwickelt daraus passende Lösungen für mittelständische Unternehmen.

Mit dieser **Working-Paper-Reihe** geben wir Einblicke in die laufende Forschung der Wissenschaftler:innen des Kompetenzzentrums und möchten gleichzeitig einen Beitrag zur Diskussion rund um aktuelle Themen aus den Feldern Künstliche Intelligenz und Arbeitsforschung leisten.

ÜBER DIE AUTOR:INNEN

Alexander Kuhn
ist wissenschaftlicher Mitarbeiter am Institut für industrielle Informationstechnik (IniT) der Technischen Hochschule OWL und nebenberuflicher IT-Berater . Fokus seiner Tätigkeit liegt im Bereich der proaktiven Ausgestaltung informationsintensiver Mensch-Maschinen-Interaktionen und des Prozessmanagements.

Stefan Hartmann
ist wissenschaftlicher Mitarbeiter am Fraunhofer-Institut für Entwurfstechnik Mechatronik IEM in Paderborn. Der Fokus seiner Tätigkeit liegt im Bereich digitale Transformation, Prozessmanagement sowie in der Etablierung und Implementierung von künstlicher Intelligenz im Produktionsumfeld.

ABSTRACT

Dieser Artikel untersucht methodische Ansätze zur Schaffung von Transparenz bei der Umsetzung von Künstlicher Intelligenz (KI) mit dem Schwerpunkt auf der Verbesserung der Interaktion zwischen Menschen, Technologie und Organisationen auf operativer Ebene.

Ein initialer Praxisexkurs, der sich auf die Entwicklung und Anwendung von Machine Learning (ML) und statistischen Methoden zur Vorhersage im Bestandsmanagement konzentriert, dient dabei als praktisches Beispiel. Bei der Anwendung dieser Methoden auf einen realen Bestandsmanagement-Datensatz wurden die Unterschiede verglichen und dabei festgestellt, welchen immensen Einfluss die Datenbeschaffenheit auf die konkrete Aussagekraft beziehungsweise Repräsentationsfähigkeit eines Modells hat. Dabei ist es sowohl für Fachleute als auch fachfremde Personen relevant, auf welcher Basis Entscheidungen getroffen, Entwicklungen forciert und auf diese Weise Ergebnisse erzielt werden können.

Im Mittelpunkt der Betrachtung steht das "Black Box"-Phänomen als anhaltendes Problem in der KI-Entwicklung. Das „Black Box"-Phänomen verweist auf die mangelnde Transparenz bei der Ausführung von Operationen innerhalb eines komplexen KI-Modells. Mangels Verständnisses können Misstrauen, Widerstand und gesellschaftliche Debatten hinsichtlich der Kontrolle und Nutzung von KI-Technologien entstehen.

Ziel dieser Ausarbeitung ist es, mit Hilfe von technologischen sowie methodischen Instrumenten die Transparenz von KI-Systemen zu verbessern und die Akzeptanz in Organisationen zu erhöhen. Basierend auf ersten Projektergebnissen und zusätzlichen Literaturrecherchen werden mögliche Unterstützungsformen und etablierte Verfahren identifiziert und präsentiert.

Um den Herausforderungen im Zusammenhang mit Akzeptanz in der KI-Entwicklung zu begegnen, ist es wichtig, innerhalb der Organisation umfassende Maßnahmen zu initiieren, die das Verständnis für KI verbessern. Mit Hilfe der Dimensionen Technik - Mensch - Organisation werden aus den vorgestellten Ansätzen relevante Leitfragen im Sinne einer reflektierenden Checkliste zusammengefasst. Im Zuge der kontinuierlichen Technologieentwicklung sollen dabei sowohl menschenzentrierte als auch organisatorische Betrachtungsebenen evaluiert und auf diese Weise eine partizipative Involvierung von Mitarbeitenden gefördert werden.

1 Einleitung

Die zunehmende Verbreitung von KI hat das Potenzial, die Effizienz und Genauigkeit in vielen Bereichen zu verbessern. Dennoch sind die Akzeptanz und Integration von KI-Systemen in Unternehmen oft mit Herausforderungen verbunden. Eine wichtige Ursache für diese Herausforderungen ist das sogenannte "Black Box"-Phänomen. Das "Black Box"-Phänomen bezieht sich auf die Schwierigkeit, die internen Prozesse und Funktionsweisen von komplexen KI-Modellen zu verstehen und ist daher ein anhaltendes Problem in der KI-Entwicklung. Dabei taucht in der Literatur immer wieder die Notwendigkeit auf, Modelle zu entwickeln, die nicht nur gute Leistung erbringen, sondern auch erklärt und verstanden werden können [1].

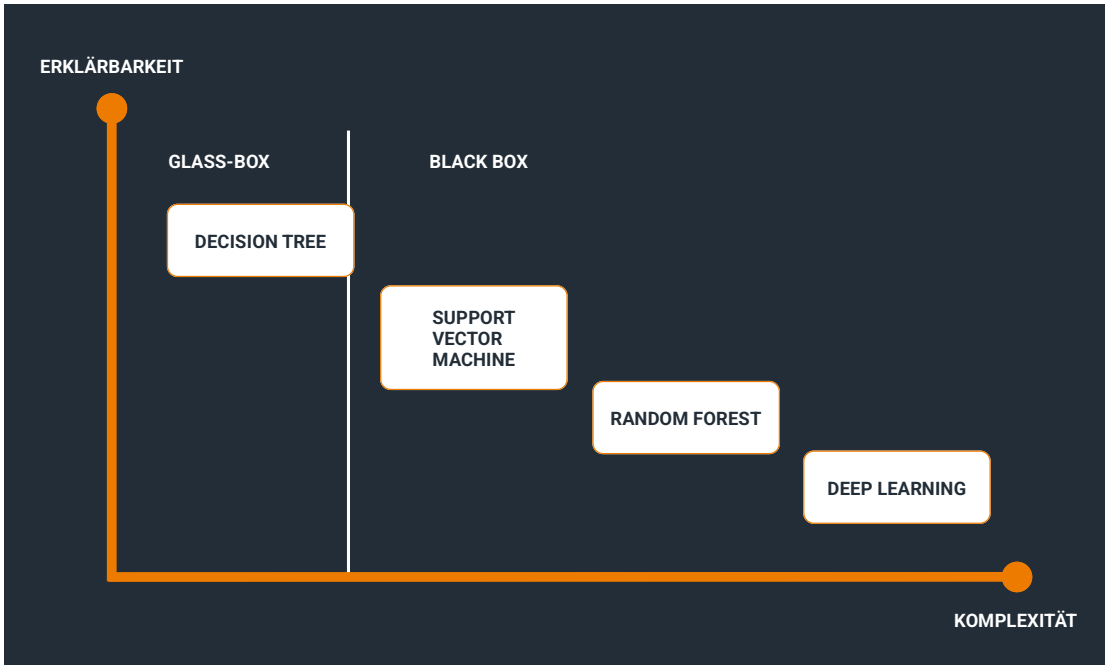

Abbildung 1: Quantitative Einordnung gängiger Machine-Learning Anwendungen [2]

Im Gegensatz zu traditionellen algorithmischen Systemen, sind KI-Modelle oft kompliziert und basieren auf maschinellem Lernen, was dazu führt, dass ihre Entscheidungen nicht immer intuitiv nachvollziehbar sind. Johner diskutiert die Herausforderungen der Interpretierbarkeit von KI-Modellen und hebt hervor, dass die komplexe Natur von KI-Modellen deren Verständlichkeit erschwert [3].

Auf individueller Ebene kann es zu Frustration führen und Misstrauen verstärken, da Benutzende unsicher sind, ob sie den Empfehlungen oder Entscheidungen von KI-Systemen vertrauen können. In Organisationen kann das Fehlen von Transparenz zu Widerstand gegenüber der Implementierung von KI führen, da Bedenken hinsichtlich der Verantwortlichkeit, Fairness und ethischen Implikationen von KI-Systemen auftreten.

Darüber hinaus verschärft diese Konstellation gesellschaftliche Debatten über die Kontrolle und den Einsatz von KI-Technologien [4].

Die grundsätzlichen Unsicherheiten, die sich aus diesem Kontext ergeben, lassen sich zudem in der Studie „Künstliche Intelligenz in der industriellen Arbeitswelt" im Zusammenhang des it's OWL Clusterprojektes Arbeitswelt.Plus zum Status Quo in der Region OstWestfalenLippe wie folgt zusammenfassen [5]:

1. **Komplexität des Themenfelds**: Die KI-Entwicklung ist mit einer Vielzahl von technischen und menschenzentrierten Herausforderungen in der Betrachtung von Arbeitsprozessen verbunden. Diese Komplexität erschwert in der Regel eine umfassende Transparenz.

2. **Heterogene KI-Eigenschaften und Aufgaben**: KI-Anwendungen weisen eine große Vielfalt an Eigenschaften und Aufgaben auf. Dies erschwert die einheitliche Darstellung und Erklärbarkeit von KI-Modellen. Eine transparente und einheitliche Kommunikation über die Funktionen und Auswirkungen der KI-Systeme ist daher eine Herausforderung.

3. **Verständnis von KI**: In einem Vergleich des Selbst- und Fremdbildes zwischen Führungs- und Mitarbeiterebene hinsichtlich der Einschätzung von KI-Kompetenzen, ergab sich eine Diskrepanz zwischen dem Selbstbild und dem Fremdbild in Bezug auf transparente Anwendung und Partizipationsmöglichkeiten, insbesondere aus Sicht der Belegschaft. Dies kann das Vertrauen in die KI-Systeme und die Bereitschaft zur aktiven Teilnahme der Mitarbeitenden beeinflussen.

4. **Weiterbildungsbedarf**: Es wird erwartet, dass der KI-Einsatz steigende Kompetenzanforderungen mit sich bringt. Es besteht ein Bedarf an Weiterbildungsmaßnahmen, besonders im Verständnis grundlegender KI-Begriffe. Das Fehlen dieser KI-Weiterbildungsangebote in vielen Unternehmen stellt dabei eine weitere Herausforderung dar.

Die Implementierung von Disziplinen der künstlichen Intelligenz stellt aufgrund ihrer Beschaffenheit Entwickler und Anwender gleichermaßen vor enorme Herausforderungen: Es handelt sich bei diesem Forschungsfeld um eine noch junge Disziplin, die durch ihre Vielfältigkeit eine enorme Komplexität aufweist. Hierbei ist nicht nur das fehlende Erfahrungswissen, wie man in bestimmten Anwendungsfällen KI einsetzt, problematisch, sondern noch mehr die Herausforderung, "die künstliche Intelligenz sichtbar zu machen".

2 Projektexkurs

Ein aktuelles Projekt im Zusammenhang des Kompetenzzentrums Arbeitswelt.Plus konzentrierte sich auf die Entwicklung und Anwendung von maschinellem Lernen (ML) und statistischen Methoden zur Vorhersage der Bestandsentwicklung.

Als Anwendungsbeispiel wurde der Absatzplanungsprozess betrachtet, der durch die Implementierung von KI-/ML-Methoden unterstützt werden soll. Die unten beschriebene Prozessmodellierung ermöglicht eine transparente Darstellung der Entwicklungen und Veränderungen für alle Stakeholder und wurde auf unterschiedlichen Abstraktionsebenen eingesetzt, um sowohl Details als auch den Gesamtüberblick darzustellen. Dabei wurden die Rollen und Kompetenzen der beteiligten Personen im Kontext der Implementierung von KI-/ML-Methoden berücksichtigt, um eine effektive Zusammenarbeit zu gewährleisten.

Es wurde festgestellt, dass ML-Techniken flexibler sind und eine Verbesserung der Vorhersagen ermöglichen. Die Ergebnisse sind jedoch oft nicht auf kleinere Datensätze übertragbar. Um die besten Vorhersagetechniken für diese Art von Datensätzen zu identifizieren, wurde eine Analyse eines realen Bestandsverwaltungsdatensatzes durchgeführt und statistische Methoden mit ML-Methoden verglichen. Die Ergebnisse zeigten, dass das LightGBM-Modell über verschiedene Szenarien und Aggregationsebenen hinweg eine konsistente Vorhersagekraft aufwies. Einfachere Methoden erwiesen sich als effektiver bei der Modellierung intermittierender oder unregelmäßiger Zeitreihen. In diesem Zusammenhang ist die Bedeutung der Wahl der Aggregationsebene hervorzuheben, da diese einen erheblichen Einfluss auf die Leistungsfähigkeit der statistischen und ML-Methoden hat. Die Studie lieferte wertvolle Erkenntnisse darüber, wie statistische und ML-Ansätze auf Datensätze unterschiedlicher Größe angewendet werden können und ob es einen Schwellenwert gibt, ab dem eine Methode besser abschneidet als die andere [6].

Neben der Entwicklung von KI-Modellen wurde im Projekt außerdem die Prozessmodellierung als Instrument zur Unterstützung der Entwicklung betrachtet. Insbesondere wurde das semantische Objektmodell (SOM) eingesetzt, um komplexe Projektierungen, wie die Implementierung von ML-/KI-Methoden, strukturiert abbilden zu können. Das SOM ermöglichte die Abbildung einer Prozesslandschaft unter Einbeziehung von menschlichen Akteuren, Systemkomponenten und Informationsflüssen. Die Modellierungstechnik diente als Orientierung, um die Auswirkungen der Implementierung von ML-/KI-Methoden transparent zu machen und die Kommunikation zwischen den Stakeholdern zu erleichtern [7].

Zusammenfassend lässt sich anhand des vorliegenden Exkurses aufzeigen, dass die Förderung der Transparenz durch diverse Maßnahmen und Komponenten maßgeblich beeinflusst wird: In dem vorliegenden Fall war es zum einen die technische Nachvollziehbarkeit bzw. Vergleichbarkeit der statistischen Verfahren im Sinne des Verhaltens und zum anderen die Sichtbarkeit der etwaigen Auswirkungen auf der Prozessebene. Durch diese schematische Strukturebene über unterschiedliche granulare Ebenen wurden etwaige Veränderungen in der Mensch-Technik-Interaktion verdeutlicht und auf einer “nicht-technischen” Ebene abgebildet.

3 Schaffung von Transparenz bei der KI-Implementierung

3.1 Erklärbarkeit von KI-Modellen

Das *Fraunhofer Institut für Intelligente Analyse- und Informationssysteme (IAIS)* hat mit seinem “KI-Prüfkatalog” im Jahr 2021 bereits ein Framework für die Bewertung der Vertrauenswürdigkeit für KI-Anwendungen erstellt. Es identifiziert dort sechs Dimensionen der Vertrauenswürdigkeit, die es im Kontext einer KI-Implementierung zu prüfen gilt [8]:

1. **Fairness**, also die Vermeidung von Diskriminierung.

2. **Autonomie und Kontrolle**, also die Wahrung des Gleichgewichts zwischen Autonomie der Anwendung und Autonomie der Nutzenden unter Sicherstellung des Vorrangs menschlicher Handlung.

3. **Transparenz**, also z. B. die technische Erklärbarkeit eines Modells.

4. **Verlässlichkeit**, also z.B. die Korrektheit von Ergebnissen oder der Immunität gegenüber manipulierten Eingaben.

5. **Sicherheit**, sowohl aus Perspektive des Schutzes vor Funktionsstörungen als auch aus Perspektive der IT-Sicherheit.

6. **Datenschutz**, also die Konformität mit allen geltenden Bestimmungen zum Schutze der Privatsphäre.

In diesem Kontext soll dabei aber der Fokus auf die Erklärbarkeit von KI-Modellen gelegt werden. Die Erklärbarkeit von KI-Modellen ist, wie eingangs beschrieben, von entscheidender Bedeutung, um das Vertrauen in KI-Systeme zu stärken und potenzielle Risiken zu minimieren. Die Forschung und Entwicklung im Bereich der Erklärbarkeit von

KI-Modellen hat in den letzten Jahren stark zugenommen. Es gibt eine Vielzahl von Ansätzen und Techniken, die darauf abzielen, KI-Modelle verständlicher und nachvollziehbarer zu machen.

Eine wichtige Methode zur Verbesserung der Erklärbarkeit von KI-Modellen besteht darin, **Feature Importance-Analysen** einzusetzen. Diese Analysen ermöglichen es, die Bedeutung einzelner Merkmale oder Variablen für die Vorhersagen des Modells zu ermitteln. Dadurch können Entscheidungen und Vorhersagen des Modells “sichtbarer” gemacht werden. Eine renommierte Quelle in diesem Bereich ist die Arbeit von Ribeiro et al. (2016), in der die Methode **"LIME"** (Local Interpretable Model-Agnostic Explanations) eingeführt wurde. LIME basiert auf Feature Importance-Analysen und ermöglicht eine interpretierbare Darstellung der Entscheidungsgrundlagen von KI-Modellen [9].

Darüber hinaus ist die Modellvisualisierung ein vielversprechender Ansatz, um die Erklärbarkeit von KI-Modellen zu verbessern. Durch den Einsatz von visuellen Darstellungen können komplexe KI-Modelle und ihre Funktionsweise besser verdeutlicht werden. Hierfür bieten Arbeiten wie Strobelt et al. (2018) wertvolle Einblicke in fortschrittliche Visualisierungstechniken für KI-Modelle. Diese Techniken ermöglichen es den Benutzenden, die Entscheidungsprozesse des Modells auf intuitivere Weise zu erfassen und potenzielle Muster oder Abhängigkeiten zu erkennen [10].

Es bestehen im Forschungsgebiet dahingehend verschiedene Ansätze und Techniken. Anhand Feature Importance-Analysen und Modellvisualisierungen können KI-Modelle durch die strukturelle Abbildung der Modellgrundlage besser verständlich gemacht werden.

3.2 Kommunikation und Schulung der Mitarbeitenden

Kommunikation und Schulungen der Mitarbeitenden spielen eine zentrale Rolle, um das KI-Verständnis innerhalb einer Organisation zu fördern. Es ist wichtig, dass die Mitarbeitenden über das Potenzial, die Grenzen und die Auswirkungen von KI informiert und sensibilisiert sind, um eine effektive Zusammenarbeit mit den Systemen zu gewährleisten.

Die Wissensvermittlung über KI-Technologien und ihre Auswirkungen ist von entscheidender Bedeutung. Hierbei können Workshops eingesetzt werden, um den Mitarbeitenden das notwendige Hintergrundwissen zu vermitteln. Solche Formate können unterschiedliche Aspekte abdecken; hierunter fallen beispielsweise die Grundlagen der KI, die Funktionsweise von spezifischen Modellen oder die ethischen Implikationen des KI-Einsatzes. Es ist wichtig, dass diese Schulungen auf die

individuellen Bedürfnisse und Vorkenntnisse der Mitarbeitenden zugeschnitten sind, um eine bestmögliche Lernumgebung zu schaffen. Quellen wie Zhang et al. (2020) haben sich mit der Gestaltung von Schulungsprogrammen für KI-Technologien auseinandergesetzt und Best Practices identifiziert [11].

Darüber hinaus spielt die Einbindung der Stakeholder, einschließlich der Mitarbeitenden, eine wesentliche Rolle bei der Schaffung von Transparenz. Transparenz kann dabei in der Dimension “Transparenz in der Implementierungsstrategie” und in der Dimension “Strategischer Nutzen von KI”, z. B. der Anpassung an den digitalen Wandel, gedacht werden. Die Mitarbeitenden sollten nicht nur informiert, sondern auch aktiv in den Implementierungsprozess von KI-Systemen einbezogen werden. Dies fördert den Austausch von Informationen, ermöglicht es den Mitarbeitenden, ihre Bedenken und Anregungen zu äußern, und trägt dazu bei, ein umfassenderes Verständnis der internen Herausforderungen und strategischen Chancen zu entwickeln. Quellen wie Dignum et al. (2019) betonen die Bedeutung einer partizipativen Gestaltung von KI-Systemen, bei der die Mitarbeitenden frühzeitig in den Entwicklungsprozess einbezogen werden [12].

4 Methoden zur Verbesserung des Zusammenspiels von Organisation, Mensch und Technik

4.1 Partizipative Gestaltung und Mitarbeitendenbeteiligung

Die partizipative Gestaltung und Beteiligung der Mitarbeitenden sind entscheidende Aspekte, um das Zusammenspiel von Organisation, Mensch und Technik zu verbessern. Es geht darum, die Mitarbeitenden aktiv in den Implementierungsprozess von KI-Systemen einzubeziehen und ihnen eine Stimme zu geben.

Bewährte Beispiele für die Einbindung von Mitarbeitenden sind unter anderem **Design Thinking** oder **Co-Creation-Workshops**. Sie ermöglichen es den Mitarbeitenden, ihre Bedürfnisse, Ideen und Perspektiven einzubringen und aktiv am Gestaltungsprozess teilzuhaben. Durch den gemeinsamen Austausch können innovative Lösungsansätze entwickelt werden, die den Anforderungen der Mitarbeitenden gerecht werden [13].

Ein weiteres Beispiel im Zusammenhang mit komplexen Informationsprojekten stellt das Konzept des **"Human-in-the-Loop"** dar, bei dem menschliches Fachwissen und Entscheidungsfindungen eng mit KI-Systemen verknüpft werden. Hierbei werden die Mitarbeitenden in Echtzeit in den Entscheidungsprozess einbezogen und haben die Möglichkeit, die Ergebnisse der KI zu überprüfen, zu korrigieren oder zu verbessern. Dieses iterative Feedback ermöglicht eine kontinuierliche Verbesserung der KI-Modelle und fördert das Vertrauen der Mitarbeitenden in die Technologie [14].

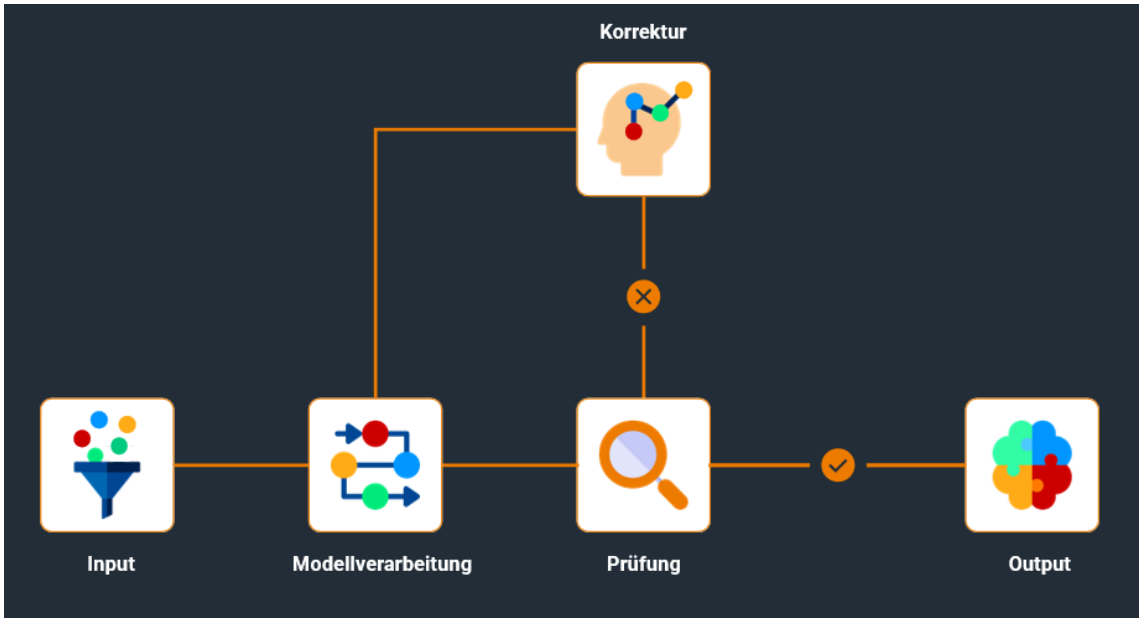

Abbildung 2: Konzept des "Human-in-the-Loop" [15]

Auf diese Weise können zum einen mögliche Bedenken und Herausforderungen frühzeitig erkannt und adressiert werden, zum anderen führt die Beteiligung der Mitarbeitenden zu einer höheren Identifikation und Akzeptanz und einem besseren Verständnis der Technologie innerhalb der Organisation.

4.2 Evaluierung und Validierung von KI-Modellen

Eine zentrale Methode zur Evaluierung von KI-Modellen besteht in der Verwendung von Performance-Metriken. Diese Metriken ermöglichen die Messung der Modellleistung anhand vordefinierter Kriterien. Beispiele für solche Metriken sind Genauigkeit, Präzision, Recall und F1-Score, die bei der Bewertung von Klassifikationsmodellen eingesetzt werden können. Die Wahl der geeigneten Metriken hängt von der spezifischen Aufgabe und den Zielen des KI-Modells ab [16].

Neben der Verwendung von Metriken sind Testszenarien ein weiteres Instrument zur Evaluierung von KI-Modellen. Dabei werden realistische Situationen geschaffen, in denen das Modell seine Fähigkeiten unter Beweis stellen muss. Diese Szenarien können sowohl simuliert als auch in einer kontrollierten Umgebung durchgeführt werden. Durch das Testen des Modells in verschiedenen Szenarien können Schwachstellen und Verbesserungspotenziale identifiziert werden. Dabei sollte zum einen die Erklärbarkeit auf lokaler Ebene, also welche Operationen zu einem konkreten Ergebnis in einem konkreten Fall geführt haben, und zum anderen die Erklärbarkeit auf globaler Ebene, also die Beschreibung der grundlegenden Treiber und Zusammenhänge eines Modells, berücksichtigt werden.[17]

Eine umfassende Dokumentation spielt ebenfalls eine entscheidende Rolle. Sie ermöglicht eine transparente Darstellung der Funktionsweise des Modells, der verwendeten Daten, der angewendeten Methoden und der erzielten Ergebnisse. Eine gute Dokumentation erleichtert nicht nur die Evaluierung und Validierung des Modells, sondern auch die Reproduzierbarkeit und den Austausch von Wissen innerhalb der Organisation.

Im Hinblick auf die Schaffung einer ganzheitlichen KI-Implementierung gibt es einige Ansätze, die zentrale Erkenntnisse und bewährte Verfahren liefern. Zwei solcher Fallstudien sind das Projekt "AI@Work" von der Stanford University und Microsoft Research sowie das Projekt "AI Fairness 360" von IBM Research.

Das Projekt "AI@Work" konzentriert sich auf die partizipative Gestaltung von KI-Systemen in verschiedenen Arbeitsumgebungen. In dieser Fallstudie werden Mitarbeitende aktiv in den Entwicklungs- und Implementierungsprozess einbezogen. Ihre Erfahrungen, Bedenken und Ideen werden berücksichtigt, um KI-Systeme so zu gestalten, dass sie den Bedürfnissen und Anforderungen der Mitarbeitenden entsprechen. Durch diesen partizipativen Ansatz wird das Verständnis für KI-Systeme gefördert und das Vertrauen der Mitarbeitenden gestärkt.

Das Projekt "AI Fairness 360" von IBM Research ist eine offene Softwarebibliothek, die Werkzeuge und Methoden zur Bewertung und Verbesserung der Fairness von KI-Modellen bereitstellt. Die Bibliothek ermöglicht es Organisationen, ihre KI-Systeme auf potenzielle Vorurteile und Diskriminierung zu überprüfen und entsprechende Maßnahmen zu ergreifen. Durch systematische Evaluierung und Verbesserung von KI-Modellen im Hinblick auf Fairness können Organisationen das Vertrauen in KI-Systeme stärken und die Akzeptanz innerhalb der Organisation fördern.

5 Zusammenfassende Handlungsempfehlungen

Durch die Integration partizipativer Gestaltungs- und Evaluierungsmethoden in den Entwicklungsprozess können Organisationen die Zusammenarbeit von Organisation, Mensch und Technikverbessern - folgende Best Practices und Handlungsempfehlungen lassen sich hieraus ableiten:

A. **Einbindung der Mitarbeitenden durch partizipative Gestaltung:** Es ist unabdingbar, die Mitarbeitenden frühzeitig in den Implementierungsprozess von KI-Systemen einzubeziehen. Dies kann beispielsweise durch partizipative Formate wie Workshops, Feedback-Sitzungen und iterative Designprozesse geschehen. Auf diese Weise gelingt eine stärkere Identifikation durch Involvierung, damit wiederum die

Steigerung der Motivation und letztendlich auch die Förderung von Verständnis und Vertrauen im Umgang mit der KI-Technologie.

B. **Evaluierung und Verbesserung der Modelle:** Die systematische Evaluierung von KI-Modellen hinsichtlich ihrer Leistung und Fairness ist essenziell. Performance-Metriken, Testszenarien und Dokumentation helfen dabei, Vorurteile, Diskriminierung und unerwünschtes Verhalten der Modelle zu erkennen und zu adressieren. Durch kontinuierliche Verbesserungen wird die Qualität und Verlässlichkeit der KI-Systeme gewährleistet.

C. **Einsatz von Tools und Methoden:** Die Nutzung von Tools und Methoden zur Interpretierbarkeit, Feature Importance-Analysen und Modellvisualisierung ermöglicht ein besseres Verständnis der KI-Modelle und ihrer Entscheidungsfindung. Durch die Anwendung von Explainable Artificial Intelligence (XAI)-Techniken können Entscheidungen nachvollzogen und verständlich gemacht werden.

D. **Schulungen und Wissensvermittlung:** Die Schulung der Mitarbeitenden im Umgang mit KI-Technologien und deren Auswirkungen ist von großer Bedeutung. Schulungen sollten nicht nur technische Aspekte abdecken, sondern auch ethische Fragestellungen, Datenschutz und die richtige Interpretation von KI-Ergebnissen behandeln. Dadurch können die Mitarbeitenden Schlüsselkompetenzen im Umgang mit KI erwerben und gewinnbringend im Arbeitsalltag einsetzen.

E. **Dokumentation und Aufbereitung**: Eine umfassende Dokumentation der Implementierungsschritte, Datenquellen, verwendeten Algorithmen und Entscheidungsprozesse ist unabdingbar. Transparente Dokumentation ermöglicht eine nachvollziehbare Überprüfung der KI-Systeme und fördert das Vertrauen sowohl intern als auch extern.

Abbildung 3: Kerndimensionen der partizipativen KI-Entwicklung

6 Ausblick

Die Schaffung von Transparenz bei der KI-Implementierung erfordert das Zusammenspiel verschiedener Maßnahmen und Ansätze. Die Erklärbarkeit von KI-Modellen, die Kommunikation und Schulung der Mitarbeitenden, die Evaluierung und Validierung von KI-Modellen, sowie die partizipative Gestaltung und Mitarbeitendenbeteiligung sind zentrale Aspekte, die zur Verbesserung der Transparenz beitragen. Beispiele wie "AI@Work" und "AI Fairness 360" liefern wertvolle Einblicke und zeigen, wie Unternehmen diese Aspekte erfolgreich umsetzen können.

Bei den vorgestellten Ansätzen gibt es jedoch auch Limitationen. Die Erklärbarkeit von komplexen KI-Modellen bleibt insbesondere bei tiefen neuronalen Netzwerken eine Herausforderung. Zudem gibt es offene Fragen hinsichtlich der Balance zwischen Erklärbarkeit und Leistungsfähigkeit der Modelle und Gewährleistung von Fairness und Diskriminierungsfreiheit. Weitere Forschung ist erforderlich, um diese Fragen zu adressieren und effektive Lösungen zu finden. Erste Bemühungen, KI in einem gesetzlichen Rahmen zu regulieren, startet aktuell das Europäische Parlament und adressiert dabei u. a. auch die Fragestellungen der Transparenz, Nachvollziehbarkeit und Diskriminierungsvermeidung [18]. In diesem Zusammenhang bleibt abzuwarten, inwiefern diese Bestimmungen klare und einfach umsetzbare Leitlinien schaffen, die den Unternehmen bei ihren Bestrebungen zur Verbesserung der Transparenz helfen können.

Die ganzheitliche Betrachtung der KI-Implementierung im Sinne Organisation, Mensch und Technik wird in Zukunft weiter an Bedeutung gewinnen. Fortschritte in den Bereichen Ethik und Fairness werden auch dazu beitragen, Vertrauen und Akzeptanz in KI-Systeme zu stärken. Zudem werden die kontinuierliche Evaluierung und Verbesserung von KI-Modellen und die Schulung der Mitarbeitenden eine immer wichtigere Rolle spielen. Die Einbindung von Share- und Stakeholdern und die Berücksichtigung von gesellschaftlichen Anliegen werden zu einer verantwortungsvollen und ethisch verbesserten Nutzung von KI beitragen.

Dabei ist es unabdingbar, dass sich Organisationen frühzeitig mit den Dimensionen Organisation, Mensch und Technik im Hinblick auf die nachhaltige KI-Umsetzung auseinandersetzen und kontinuierlich reflektierend anwenden. Ansonsten besteht die Gefahr, dass die notwendige Nachvollziehbarkeit leidet und zu einer Verschärfung der “Black-Box” führt.

7 Literaturverzeichnis

[1] Käde, L., von Maltzan S. (2020). "Die Erklärbarkeit von Künstlicher Intelligenz (KI): Entmystifizierung der Black Box und Chancen für das Recht" *Computer und Recht*, vol. 36, no. 1, 2020 (pp. 66-72).

[2] ppi AG (2023). https://www.ppi.de/banken/think-banking/ki-tatsaechlich-mehrwertig-einsetzen/explainable-ai/ [Abgerufen am: 04.07.2023 | 14:05].

[3] Johner, C. (2022). Interpretierbarkeit von KI: Blick in die Blackbox des maschinellen Lernens.

[4] Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. Big Data & Society, 3(1), 2053951715622512.

[5] Papenkordt, J; Gabriel, S.; Thommes, K.; Dumitrescu, R. (2022). Künstliche Intelligenz in der industriellen Arbeitswelt - Studie zum Status Quo in der Region OstWestfalenLippe.

[6] Pfeifer, A.; Brandt, H.; Lohweg, V. (2023). A Comparison of Statistical and Machine Learning Approaches for Time Series Forecasting in a Demand Management Scenario, In: IEEE 21st International Conference on Industrial Informatics (INDIN).

[7] Kuhn, A. & Dyck, F. (2022). Menschenzentrierte Arbeitsprozessmodellierung - Kommunikationsmedium für Veränderungen von Mensch-Maschine-Interaktionen im Zuge der Implementierung von KI.

[8] Poretschkin, M.; Schmitz, A.; Akila, M.; Adilova, L.; Becker, D.; Cremers, A. B.; Hecker, D.; Houben, S.; Mock, M.; Rosenzweig, J.; Sicking, J.; Schulz, E.; Voß, A.; Wrobel, S. (2021). Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz (KI-Prüfkatalog): Fraunhofer IAIS. Online verfügbar unter https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/ki-pruefkatalog/202107_KI-Pruefkatalog.pdf.

[9] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144).

[10] Strobelt, H., Gehrmann, S., Pfister, H., & Rush, A. M. (2018). LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks. IEEE Transactions on Visualization and Computer Graphics, 24(1), 667-676.

[11] Zhang, L., Cui, P., & Wang, W. (2020). Challenges and Opportunities: Designing AI Training Programs for the General Public. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 229-238).

[12] Dignum, V., Noriega, P., Rodriguez-Aguilar, J. A., & Parsons, S. (2019). Governance and Accountability of Artificial Intelligence Systems. AI & Society, 34(3), 547-557.

[13] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. Springer.

[14] Huber, M. (2022). KI unter Kontrolle. Interaktiv Online. Abgerufen am 13.Juli 2023, von https://interaktiv.ipa.fraunhofer.de/kuenstliche-intelligenz-fuer-die-produktion/ki-unter-kontrolle/.

[15] Klippa App B.V. (2022). https://www.klippa.com/en/blog/information/human-in-the-loop/ [Abgerufen am: 01.06.2023 | 16:13].

[16] Väänänen-Vainio-Mattila, K. (2008). User involvement in the early phases of innovation process. International Journal of Product Development, 6(1-2), 79-95.

[17] Rader, E., Goldberg, Y., & Perkins, T. (2018). Weakly supervised extraction of computer vision benchmarks from the web. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 424-440).

[18] Europäisches Parlament (2023). KI-Gesetz: erste Regulierung der künstlichen Intelligenz. Abgerufen am 18.07.2023, von https://www.europarl.europa.eu/news/de/headlines/society/20230601STO93804/ki-gesetz-erste-regulierung-der-kunstlichen-intelligenz?&at_campaign=20226-Digital&at_medium=Google_Ads&at_platform=Search&at_creation=RSA&at_goal=TR_G&at_advertiser=Webcomm&at_audience=ki-gesetz&at_topic=Artificial_intelligence_Act&at_location=DE&gclid=Cj0KCQjw8NilBhDOARIsAHzpbLApRKdTmpFDuQIyWS0CKWIFRXclZo2kZmn6UTYvhVgmz5OCafKu0loaAqd8EALw_wcB.

IMPRESSUM

Verantwortlich für den Inhalt
Alexander Kuhn, Institut für Industrielle Informationstechnik, Technische Hochschule OWL
Stefan Hartmann, Fraunhofer-Institut für Entwurfstechnik Mechatronik IEM

Fotos/Abbildungen
Titel: unsplash.com | Nicolas Arnold
S. 2: InIT, Fraunhofer IEM

Gestaltung & Redaktion
Salome Leßmann
it's OWL Clustermanagement GmbH

Empfohlene Zitierweise
Kuhn, A.; Hartmann, S. (2023): Das "Black-Box-Phänomen" in der KI-Entwicklung - Methodische Ansätze zur Schaffung von Transparenz und der Verbesserung des Zusammenspiels von Mensch, Technik und Organisation. Working-Paper-Reihe des Kompetenzzentrums Arbeitswelt.Plus, Paderborn, Nr. 8, https://doi.org/10.55594/SFAH4426

Erscheinung
09/2023

Möchten Sie mehr über die Forschungsarbeit im Kompetenzzentrum Arbeitswelt.Plus erfahren? Auf unserer Website finden Sie detaillierte Informationen zu allen Forschungsschwerpunkten.

Kompetenzzentrum Arbeitswelt.Plus
c/o it's OWL Clustermanagement GmbH
Zukunftsmeile 2
33012 Paderborn
www.arbeitswelt.plus

Kompetenzzentren Arbeitsforschung

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Dieses Forschungs- und Entwicklungsprojekt wird durch das Bundesministerium für Bildung und Forschung (BMBF) im Programm „Zukunft der Wertschöpfung – Forschung zu Produktion, Dienstleistung und Arbeit" gefördert und vom Projektträger Karlsruhe (PTKA) betreut. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei der Autorin / beim Autor.